

# Generalization of Clustering Agreements and Distances for Overlapping Clusters and Network Communities

Reihaneh Rabbany · Osmar R. Zaiane

Received: date / Accepted: date

**Abstract** A measure of distance between two clusterings has important applications, including clustering validation and ensemble clustering. Generally, such distance measure provides navigation through the space of possible clusterings. Mostly used in cluster validation, a normalized clustering distance, a.k.a. agreement measure, compares a given clustering result against the ground-truth clustering. Clustering agreement measures are often classified into two families of pair-counting and information theoretic measures, with the widely-used representatives of Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), respectively. This paper sheds light on the relation between these two families through a generalization. It further presents an alternative algebraic formulation for these agreement measures which incorporates an intuitive clustering distance, which is defined based on the analogous between cluster overlaps and co-memberships of nodes in clusters. Unlike the original measures, it is easily extendable for different cases, including overlapping clusters and clusters of inter-related data for complex networks. These two extensions are, in particular, important in the context of finding clusters in social and information networks, a.k.a communities.

**Keywords** Clustering Agreement; Cluster Evaluation; Cluster Validation; Network Clusters; Community Detection; Overlapping Clusters

## 1 Introduction

A cluster distance, accordance, similarity, or divergence has different applications. *Cluster validation* is the most common usage of cluster distance measures. In particular, in external evaluation, a clustering algorithm is validated on a set of benchmark datasets by comparing the similarity of its results against the ground-truth clusterings. Another notable application is *ensemble, or consensus Clustering*, where results of different clustering algorithms on the same dataset are aggregated. A notion of distance between alternative clusterings is used in modeling and formulating this aggregation, i.e. to find a clustering that has the minimum average distance

---

Department of Computing Science, University of Alberta  
E-mail: {rabbanyk,zaiane}@ualberta.ca

to the alternative clusterings<sup>1</sup>. Another closely related application is multi-view clustering (Cui et al, 2007), where the objective is to find different clusterings of the same dataset, which are usually in different sub-spaces of the data, and could represent different views of that dataset. In the same context, one might be interested to find the sub-spaces that result in different/similar clusterings.

Clustering distance measures are well-studied and widely-used in cluster validation, where a normalized distance measure is used to average the performance of an algorithm over different datasets, and to compare different algorithms. Some of the most widely used clustering agreement measures are: Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Variation of Information (VI).

In this paper, we first study the well-known clustering agreement measures, which are classified into two families of pair counting and information theoretic measures. Then we highlight the relation between these two families by presenting a generalized formula that covers both. Next, we elaborate on the limitations of these measures in handling inter-related data-points, and also overlapping clusters. These two limitations are in particular problematic when measuring distance between clusterings in the context of information networks.

Networks encode the relationship between data-points, and clusters on a real network are known to be overlapping. Many methods for network clustering, a.k.a. community mining, have been proposed in recent years; the reader could refer to Fortunato (2010) for a survey. In the evaluation and comparison of these algorithms, often the classical clustering agreement measures, mostly NMI, are applied. Here, we discuss the effect of neglecting relations between data points, e.g. edges in networks, in measuring communities distance, and derive extensions of our generalized formula to incorporate such relationships.

We further discuss the difficulty of extending the current contingency (a.k.a. overlap, or confusion) based formulation for the general cases of overlapping clusters. We tackle this by presenting an alternative algebraic formulation for a clustering distance, based on the analogous relationship of cluster overlaps and co-memberships of nodes. From the proposed algebraic formulation we could derive the original formulations, and we could also easily derive new forms that are appropriate for the cases of overlapping clusters, and also network clusters.

## 2 Clustering Agreement Measures: Short Survey

Consider a dataset  $D$  consisting of  $n$  data items,  $D = \{d_1, d_2, d_3 \dots d_n\}$ . A partitioning  $U$  partitions  $D$  into  $k$  mutually disjoint subsets,  $U = \{U_1, U_2 \dots U_k\}$ ; where  $D = \cup_{i=1}^k U_i$  and  $U_i \cap U_j = \emptyset \forall i \neq j$ . There are several measures defined to examine the similarity, a.k.a agreement, between two partitioning of the same dataset. More formally, let  $V$  denote another partitioning of the dataset  $D$ ,  $V = \{V_1, V_2 \dots V_r\}$ . Clustering agreement measures are originally introduced based on counting the pairs of data items that are in the same/different partition in  $U$  and  $V$ . Each pair  $(d_i, d_j)$  of data items is classified into one of four groups based on their co-memberships in  $U$  and  $V$ ; which results in the following pair-counts.

---

<sup>1</sup> Refer to Aggarwal and Reddy (2014), Chapter 23 on clustering validation measures (in particular the section on external clustering validation measures); and Chapter 22 on cluster ensembles (in particular the section on measuring similarity between clustering solutions).

|                  | Same in $V$   | Different in $V$ |
|------------------|---------------|------------------|
| Same in $U$      | $M_{11} = TP$ | $M_{10} = FP$    |
| Different in $U$ | $M_{01} = FN$ | $M_{00} = TN$    |

Here,  $M_{11}/M_{00}$  counts the number of pairs that are in the same/different partitions in both  $U$  and  $V$ .  $M_{10}/M_{01}$  sums up those that belong to the same/different partitions in  $U$  but are in different same/partitions according to  $V$ . Note that  $M_{11} + M_{00} + M_{10} + M_{01} = \binom{n}{2}$ . When one of these partitionings, for instance  $V$ , is the true partitioning i.e. the ground-truth, these could also be referred to as the true/false positive/negative scores, denoted by TP, FP, TN, and FN in the table<sup>2</sup>.

These pair counts are often derived using the following contingency table a.k.a. confusion table (Hubert and Arabie, 1985). The contingency table is a  $k \times r$  matrix of all the possible overlaps between each pair of clusters in  $U$  and  $V$ , where its  $ij$ th element shows the intersection of cluster  $U_i$  and  $V_j$ , i.e.  $n_{ij} = |U_i \cap V_j|$ .

|               | $V_1$         | $V_2$         | $\dots$  | $V_r$         | marginal sums |
|---------------|---------------|---------------|----------|---------------|---------------|
| $U_1$         | $n_{11}$      | $n_{12}$      | $\dots$  | $n_{1r}$      | $n_{1\cdot}$  |
| $U_2$         | $n_{21}$      | $n_{22}$      | $\dots$  | $n_{2r}$      | $n_{2\cdot}$  |
| $\vdots$      | $\vdots$      | $\vdots$      | $\ddots$ | $\vdots$      | $\vdots$      |
| $U_k$         | $n_{k1}$      | $n_{k2}$      | $\dots$  | $n_{kr}$      | $n_{k\cdot}$  |
| marginal sums | $n_{\cdot 1}$ | $n_{\cdot 2}$ | $\dots$  | $n_{\cdot r}$ | $n$           |

The last row and column show the marginal sums of  $n_{i\cdot} = \sum_j n_{ij}$ , and  $n_{\cdot j} = \sum_i n_{ij}$ , where in this case of disjoint clusters we also have  $n_{i\cdot} = |U_i|$ , and  $n_{\cdot j} = |V_j|$ . The pair counts can then be computed using the following formulae.

$$M_{10} = \sum_{i=1}^k \binom{n_{i\cdot}}{2} - \sum_{i=1}^k \sum_{j=1}^r \binom{n_{ij}}{2}, \quad M_{01} = \sum_{j=1}^r \binom{n_{\cdot j}}{2} - \sum_{i=1}^k \sum_{j=1}^r \binom{n_{ij}}{2}$$

$$M_{11} = \sum_{i=1}^k \sum_{j=1}^r \binom{n_{ij}}{2}, \quad M_{00} = \binom{n}{2} + \sum_{i=1}^k \sum_{j=1}^r \binom{n_{ij}}{2} - \sum_{i=1}^k \binom{n_{i\cdot}}{2} - \sum_{j=1}^r \binom{n_{\cdot j}}{2}$$

These *pair counts* have been used to define a variety of different clustering agreement measures (Manning et al, 2008). Here, we briefly explain the most common measures; the reader can refer to Albatineh et al (2006) for a complete survey.

Considering co-membership of data points in the same or different clusters as a binary variable, *Jaccard* agreement between clustering  $U$  and  $V$  can be defined as  $J = TP/(FP + FN + TP) = M_{11}/(M_{01} + M_{10} + M_{11})$ . *Rand Index* is defined similarly to Jaccard, but it also values pairs that belong to different clusters in both partitionings, i.e. true negatives:  $RI = (M_{11} + M_{00})/(M_{11} + M_{01} + M_{10} + M_{00})$ , which gives:

$$RI = 1 + \frac{1}{n^2 - n} \left( 2 \sum_{i=1}^k \sum_{j=1}^r n_{ij}^2 - \left( \sum_{i=1}^k n_{i\cdot}^2 + \sum_{j=1}^r n_{\cdot j}^2 \right) \right) \quad (1)$$

The *Mirkin Index* is a transformation of Rand Index, defined as  $n(n-1)(RI-1)$ , which is equivalent to  $RI$  when comparing partitionings of the same dataset (Wu et al, 2009). *F-measure* is a weighted mean of the precision ( $P$ ) and recall ( $R$ ),  $F_\beta = \frac{(\beta^2+1)PR}{\beta^2P+R}$  where  $P = M_{11}/(M_{11} + M_{10})$  and  $R = M_{11}/(M_{11} + M_{01})$ . The parameter

<sup>2</sup> Also denoted by  $a, b, c, d$  letters for the notational convenience in some literature.

$\beta$  indicates how much recall is more important than precision. The two common values for  $\beta$  are 2 and .5; the former weighs recall higher than precision while the latter favours the precision more.

There is also a family of *information theoretic* based measures. These measures consider the overlaps between clusters in  $U$  and  $V$ , as a joint distribution of two random variables, i.e. the cluster memberships in  $U$  and  $V$ . The entropy of cluster  $U$ ,  $H(U)$ , the joint entropy of  $U$  and  $V$ ,  $H(U, V)$ , their mutual information,  $I(U, V)$ , and their *Variation of Information* (Meilă, 2007),  $VI(U, V)$  are then defined as:

$$\begin{aligned} H(U) &= - \sum_{i=1}^k \frac{n_{i.}}{n} \log\left(\frac{n_{i.}}{n}\right), & H(V) &= - \sum_{j=1}^r \frac{n_{.j}}{n} \log\left(\frac{n_{.j}}{n}\right) \\ H(U, V) &= - \sum_{i=1}^k \sum_{j=1}^r \frac{n_{ij}}{n} \log\left(\frac{n_{ij}}{n}\right), & I(U, V) &= \sum_{i=1}^k \sum_{j=1}^r \frac{n_{ij}}{n} \log\left(\frac{n_{ij}/n}{n_{i.}n_{.j}/n^2}\right) \\ VI(U, V) &= \sum_{i=1}^k \sum_{j=1}^r \frac{n_{ij}}{n} \log\left(\frac{n_{i.}n_{.j}/n^2}{n_{ij}^2/n^2}\right) \end{aligned} \quad (2)$$

All the pair counting measures defined here, except Mirkin, have a fixed range of  $[0, 1]$ . The above information theoretic measures, however, do not have a fixed range. For example, the mutual information ranges between  $(0, \log k]$ , and the range for variation of information is  $[0, 2 \log \max(k, r)]$  (Wu et al, 2009). Having a fixed range, i.e. being *normalized*, is a desired property for partitioning agreement indexes, since we often require to compare/average agreements over different datasets. Consequently, normalized variations of mutual information are defined (Vinh et al, 2010). The most commonly used normalization forms are:

$$NMI_{\Sigma} = \frac{2I(U, V)}{H(U) + H(V)} \quad \text{and} \quad NMI_{\sqrt{}} = \frac{I(U, V)}{\sqrt{H(U)H(V)}} \quad (3)$$

Beside having a fixed range, a clustering agreement measure should also have a *constant baseline* (Vinh et al, 2010; Hubert and Arabie, 1985). As an example, consider the case where agreement between a clustering and the ground-truth is measured as 0.7. If the baseline of the measure is not constant, it can be 0.6 in one settings and 0.2 in another, then this 0.7 value can be both a strong or a weak agreement. *Correction for chance* is adjusting a measure to have a constant (usually 0) expected value for agreements no better than random. This adjustment is calculated based on an upper bound on the measure,  $Max[M]$ , and its expected value,  $E[M]$ , as:

$$AM = \frac{M - E[M]}{Max[M] - E[M]} \quad (4)$$

The *Adjusted Rand Index* ( $ARI$ ) is proposed in (Hubert and Arabie, 1985), assuming that the contingency table is constructed randomly when the marginals are fixed, i.e. the size of the clusters in  $U$  and  $V$  are fixed. With this assumption,  $RI$  is a linear transformation of  $\sum_{i,j} \binom{n_{ij}}{2}$ , and  $E\left(\sum_{i,j} \binom{n_{ij}}{2}\right) = \sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} / \binom{n}{2}$ . Hence, adjusting  $RI$  with upper bound 1 results in the following formula:

$$ARI = \frac{\sum_{i=1}^k \sum_{j=1}^r \binom{n_{ij}}{2} - \sum_{i=1}^k \binom{n_{i.}}{2} \sum_{j=1}^r \binom{n_{.j}}{2} / \binom{n}{2}}{\frac{1}{2} \left[ \sum_{i=1}^k \binom{n_{i.}}{2} + \sum_{j=1}^r \binom{n_{.j}}{2} \right] - \sum_{i=1}^k \binom{n_{i.}}{2} \sum_{j=1}^r \binom{n_{.j}}{2} / \binom{n}{2}} \quad (5)$$

There is also an approximate formulation (Hubert and Arabie, 1985; Albatineh et al, 2006) for this expectation defined as  $E(\sum_{i,j} n_{ij}^2) = \sum_i n_{i.}^2 \sum_j n_{.j}^2 / n^2$ , which results in a slightly different formula for the *ARI*, i.e.

$$ARI' = \frac{\sum_{i=1}^k \sum_{j=1}^r n_{ij}^2 - \sum_{i=1}^k n_{i.}^2 \sum_{j=1}^r n_{.j}^2 / n^2}{\frac{1}{2} [\sum_{i=1}^k n_{i.}^2 + \sum_{j=1}^r n_{.j}^2] - \sum_{i=1}^k n_{i.}^2 \sum_{j=1}^r n_{.j}^2 / n^2} \quad (6)$$

There are several variations of pair counting agreement measures, such as Gamma, Hubert, Pearson, etc. These measures, however, become similar or even equivalent after correction for chance. More specifically, Albatineh et al (2006) show that many of these measures are linear transformations of  $\sum_{i,j} n_{ij}^2$ , i.e. each measure could be written as  $\alpha + \beta \sum_{i,j} n_{ij}^2$ , where  $\alpha$  and  $\beta$  depend on the marginal counts,  $n_{i.}$  or  $n_{.j}$ , but not on the  $n_{ij}$ . For example for the Rand Index we have:  $\alpha = 1 - \frac{1}{n(n-1)} (\sum_i n_{i.}^2 + \sum_j n_{.j}^2)$ , and  $\beta = 2/n(n-1)$ . They further prove that these measures become equivalent if their  $\frac{1-\alpha}{\beta}$  ratio is the same, since their corrected for chance formula will all be as:

$$\frac{\sum_{i,j} n_{ij}^2 - E(\sum_{i,j} n_{ij}^2)}{1 - \alpha/\beta - E(\sum_{i,j} n_{ij}^2)}$$

Warrens (2008a) extended these results and included the inter-rater reliability indices from *statistics*. Using the  $2 \times 2$  pair counting table, he has shown that all the pair counting clustering agreement measures after correction for chance become equivalent to one of the statistical inter-rater agreement indices. The well-studied *inter-rater agreement indices* in statistics are defined to measure the agreement between different coders, rankers, or judges on categorizing the same data. Examples are the goodness of fit: chi-square test, the likelihood chi-square, kappa measure of agreement, Fisher's exact test, Krippendorff's alpha, etc. (see test 16 in (Cortina-Borja, 2012)). These statistical tests are also defined based on the contingency table which displays the multivariate frequency distribution of the (categorical) variables. Specifically, *Cohen's kappa* is one the most widely used inter-rater agreement index; a chance corrected index of association defined for accessing the agreement between two raters, who categorize data into  $k$  categories (defined as  $\kappa = \frac{\sum_{i=j}^k n_{ij} - \sum_{i=j}^k E_{ij}}{n - \sum_{i=j}^k E_{ij}}$  where  $E_{ij} = \frac{n_{i.} n_{.j}}{n}$ ). The equivalence of Cohen's kappa and the *ARI* is proved by Warrens (2008b).

Vinh et al (2009) proposed the correction for chance of the information theoretic measures, and showed that Adjusted Variation of Information (*AVI*) is equivalent to *Adjusted Mutual Information (AMI)*. They derived the expected value of the mutual information assuming the sizes of the clusters are fixed, i.e. similar to the *ARI*'s assumption on the hypergeometric model of randomness. In more details, the expected value is defined as:

$$E[I(U, V)] = \sum_{i,j} \sum_{m=\max(n_{i.}, n_{.j}-n, 1)}^{\min(n_{i.}, n_{.j})} \frac{1}{n} \log\left(\frac{nm}{n_{i.} n_{.j}}\right) \frac{n_{i.}! n_{.j}! (n-n_{i.})! (n-n_{.j})!}{n! m! (n_{i.}-m)! (n_{.j}-m)! (n-n_{i.}-n_{.j}+m)!}$$

From which,  $I$  can be adjusted for chance using Equation 4:  $AMI = \frac{I - E[I]}{Max[I] - E[I]}$ ; where  $Max[I]$  is one of upper bounds on  $I$ :

$$I(U, V) \leq \min(H(U), H(V)) \leq \sqrt{H(U)H(V)} \leq \frac{H(U)+H(V)}{2} \leq \max(H(U), H(V)) \leq H(U, V)$$

The  $AVI = AMI$  is true when the  $1/2(H(U) + H(V))$  upper bound is used in the adjustment. The formulation of  $AMI$  includes big factorials, therefore is computationally complex, and less practical when compared to the  $ARI$ .

### 3 Generalization of Clustering Agreement Measures

In this section, we highlight the connection between pair counting and information theoretic measures, through defining a generalized formula that covers both. We start by noting the relation between the Rand Index ( $RI$ ), as a representative of the pair counting measures, and the Variation of Information ( $VI$ ), as a representative for the information theoretic measures.

**Proposition 1**  *$VI$  ( $RI$ ) of two partitionings is proportional to the conditional entropies (variances) of memberships in them (see Appendix A.1 for proof), i.e.*

$$VI(U, V) = H(U|V) + H(V|U) \quad \text{and} \quad RI(U, V) \propto Var(U|V) + Var(V|U)$$

This proposition inspires defining a generalized distance for clusterings as:

**Definition 1** Generalized Clustering Distance ( $\mathcal{D}$ )

$$\mathcal{D}_{\varphi}^{\eta}(U, V) = \mathcal{D}_{\varphi}^{\eta}(U||V) + \mathcal{D}_{\varphi}^{\eta}(V||U), \quad \mathcal{D}_{\varphi}^{\eta}(U||V) = \sum_{v \in V} \left[ \varphi \left( \sum_{u \in U} \eta_{uv} \right) - \sum_{u \in U} \varphi(\eta_{uv}) \right]$$

where  $\eta_{uv}$  quantifies the similarity between the two clusters of  $u \in U$  and  $v \in V$ , i.e.  $\eta : 2^V \times 2^U \rightarrow \mathbb{R}$ ; and  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ .

**Corollary 1**  $\mathcal{D}$  is bounded if  $\varphi$  is a positive superadditive function (proof in Appendix A.2), i.e.

$$\varphi(x) \geq 0 \wedge \varphi(x+y) \geq \varphi(x) + \varphi(y) \implies 0 \leq \mathcal{D}_{\varphi}^{\eta}(U||V) \leq \varphi \left( \sum_{v \in V} \sum_{u \in U} \eta_{uv} \right)$$

Using this bound as a normalizing factor, we define:

**Definition 2** Normalized Generalized Clustering Distance ( $\mathcal{ND}$ )

$$\mathcal{ND}_{\varphi}^{\eta}(U, V) = \frac{\mathcal{D}_{\varphi}^{\eta}(U, V)}{NF(U, V)}, \quad NF(U, V) = \varphi \left( \sum_{v \in V} \sum_{u \in U} \eta_{uv} \right)$$

We can show that the following two identities hold for the proposed  $\mathcal{ND}$ .

**Identity 1** *The Variation of Information (Equation 2) derives from  $\mathcal{ND}$  if we set  $\varphi(x) = x \log x$ , and  $\eta$  as the overlap size:  $\eta_{uv} = |u \cap v|$  (proof in Appendix A.3), i.e.*

$$\mathcal{ND}_{x \log x}^{|\cap|}(U, V) \equiv \frac{VI(U, V)}{\log n}$$

**Identity 2** *The Rand Index (Equation 1) derives from  $\mathcal{ND}$  if we set  $\varphi(x) = \binom{x}{2}$ , and  $\eta$  as the overlap size (proof in Appendix A.4), i.e.*

$$\mathcal{ND}_{\binom{x}{2}}^{|\cap|}(U, V) \equiv 1 - RI(U, V)$$

Similar to the Identity 2, in the rest of this paper we consider clustering agreement ( $\mathcal{I}$ ) and normalized distance ( $\mathcal{ND}$ ) interchangeably using  $\mathcal{I} = 1 - \mathcal{ND}$ .

We further adjust the generalized distance to return the maximum of one, if  $U$  and  $V$  are independent. Assume the joint probability distribution  $P_{U,V}(u, v) = \eta_{uv} / \sum_{uv} \eta_{uv}$ , with the marginal probabilities of  $P_U(u) = \sum_v P_{U,V}(u, v) = \eta_{\cdot v} / \sum_{uv} \eta_{uv}$  and  $P_V(v) = \eta_{u \cdot} / \sum_{uv} \eta_{uv}$ . Then the independence condition for  $U$  and  $V$ ,  $P_{U,V}(u, v) = P_U(u)P_V(v)$ , translates into  $\eta_{uv} = \eta_{u \cdot} \eta_{\cdot v} / \sum_{uv} \eta_{uv}$ . On the other hand, we have  $\mathcal{D}_\varphi^\eta(U, V) = \sum_{v \in V} \varphi(\eta_{\cdot v}) + \sum_{u \in U} \varphi(\eta_{u \cdot}) - 2 \sum_{v \in V} \sum_{u \in U} \varphi(\eta_{uv})$ , hence we define:

**Definition 3** Adjusted Generalized Clustering Distance ( $\mathcal{AD}$ )

$$\mathcal{AD}_\varphi^\eta = \frac{\mathcal{D}_\varphi^\eta(U, V)}{NF(U, V)}, \quad NF = \sum_{v \in V} \varphi(\eta_{\cdot v}) + \sum_{u \in U} \varphi(\eta_{u \cdot}) - 2 \sum_{u \in U} \sum_{v \in V} \varphi\left(\frac{\eta_{\cdot v} \eta_{u \cdot}}{\sum_{u \in U} \sum_{v \in V} \eta_{uv}}\right)$$

**Identity 3** The Normalized Mutual Information (Equation 3) derives from  $\mathcal{AD}$ , if we set  $\varphi(x) = x \log x$ , and  $\eta$  as the overlap size:  $\eta_{uv} = |u \cap v|$  (proof in Appendix A.5), i.e.

$$\mathcal{AD}_{x \log x}^{|\cap|}(U, V) \equiv 1 - NMI_{sum}(U, V)$$

**Identity 4** The Adjusted Rand Index of Equation 5 and Equation 6 derive from  $\mathcal{AD}$ , if we set  $\varphi(x) = x(x-1)$  and  $\varphi(x) = x^2$  respectively, where  $\eta$  is the overlap size, (proof in Appendix A.5), i.e.

$$\mathcal{AD}_{x^2}^{|\cap|}(U, V) \equiv 1 - ARI'(U, V), \quad \mathcal{AD}_{x(x-1)}^{|\cap|}(U, V) \equiv 1 - ARI(U, V)$$

This line of generalization is similar to the works in Bergman Divergence and  $f$ -divergences. For example, the mutual information and variance are proved to be special cases of Bergman information (Banerjee et al, 2005). The (reverse) KL divergence and Pearson  $\chi^2$  are shown to be  $f$ -divergences when the generator is  $x \log x$  and  $(x-1)^2$  respectively (Nielsen and Nock, 2013). Beside this analogy, our generalized measure is different from these divergences. One could consider our proposed measure as an (adjusted normalized) conditional Bergman entropy for clusterings. This relation is however non-trivial and is out of scope of this paper.

### 3.1 Extension for Inter-related Data

All the agreement measures presented so far only consider memberships of data-points, and ignore any relations between them. Ignoring these relations is however problematic, as also mentioned by a few previous works. For example Zhou et al (2005) discuss the issue of ignoring the distances between data-points when comparing clusterings, and propose to compare clusterings using a measure that incorporates the distances between representatives of the clusters.

The extension of the clustering agreement or distance measures to incorporate the structure of the data, is in particular important when comparing *clusterings of nodes within information networks*. An information network encodes relationships between data points, and a clustering on such network forms sub-graphs. Using the original clustering agreement measures to compare there clusterings, we only consider the nodes in measuring the clustering distance. It is however relevant that one should also consider edges when comparing two sub-graphs. Figure 1 presents a clarifying example for the effect of considering or neglecting edges in comparing the network clusterings, a.k.a. communities.

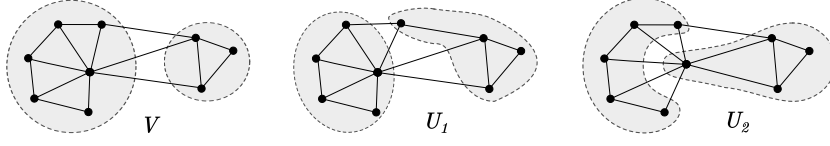


Fig. 1: Partitioning  $U_1$  and  $U_2$  of the same graph with true partitioning  $V$ . Considering only the number of nodes in the overlaps,  $U_1$  and  $U_2$  have the same contingency table with  $V$ , i.e.  $|\cap|(U_1, V) = |\cap|(U_2, V) = \{\{5, 0\}, \{1, 3\}\}$ . Therefore they have the same agreement with  $V$ , regardless of the choice of the agreement measure:  $ARI$ ,  $NMI$ , etc. However if considering the edges,  $U_1$  is more similar to the true partitioning  $V$ . This could be enforced using an alternative overlap function that incorporates edges, such as the degree weighted overlap function:  $\Sigma d(U_1, V) = \{\{18, 0\}, \{3, 9\}\}$  and  $\Sigma d(U_2, V) = \{\{14, 0\}, \{7, 9\}\}$ ; or the edge based variation:  $\xi(U_1, V) = \{\{7, 0\}, \{0, 3\}\}$  and  $\xi(U_2, V) = \{\{4, 0\}, \{0, 3\}\}$ .

To incorporate the structure in our generalized distance measure, we can modify the overlap function  $\eta$  in Definition 1. The overlap function from which the original  $RI$  or  $VI$  derive can be written as  $|\cap| : \eta = \sum_{i \in u \cap v} 1$ . Therefore the first intuitive modification to incorporate the structure is to consider a degree weighted function as:

$$\Sigma d : \eta_{uv} = \sum_{i \in u \cap v} d_i$$

Using this  $\eta$ , well-connected nodes with higher degree weigh more in the distance. Alternatively, any other ranking criteria can be used depending on the underlying application. Another possibility is to alter  $\eta$  to directly assess the structural similarity of these sub-graphs by counting their common edges:

$$\xi : \eta_{uv} = \sum_{i, j \in u \cap v} A_{ij}$$

One can consider many other alternatives for measuring overlap of two sub-graphs based on the application at hand. We revisit and delve deeper in this topic in Section 4.1, after providing an alternative formulation for the clustering distance or agreement measures.

### 3.2 Extension for Overlapping Clusters

There are several non-trivial extensions of the clustering agreement measures for the crisp overlapping clusters (Collins and Dent, 1988; Lancichinetti et al, 2008a; Xie et al, 2013). Notably, Collins and Dent (1988) proposed the **Omega index** as a generalization of the (adjusted) rand index for non-disjoint clusters with crisp memberships. The Omega index expands the  $2 \times 2$  pair-counts table of  $U$  and  $V$ ,  $\{\{M_{00}, M_{10}\}, \{M_{01}, M_{11}\}\}$ ; so that  $M_{ij}$  counts the pair of data points that appeared together in  $i$  clusters of  $U$  and  $j$  clusters of  $V$ . Similar to the  $RI$ , trace of this matrix, i.e.  $\sum_i M_{ii}$ , is considered as the agreement index, which is further adjusted for chance using marginals of  $M$ . The Omega index reduces to the  $(A)RI$  if the clusterings are disjoint. It however has a fundamental problem as it only considers the pairs that appeared in the *exact* same number of clusters together. For example, consider a pair of data points which are in 2 clusters together in the ground-truth. The Omega agreement of a clustering that puts that pair together in 1 cluster is the same as another clustering that puts them together in no clusters. Figure 2a provides an illustrated example for such a case.



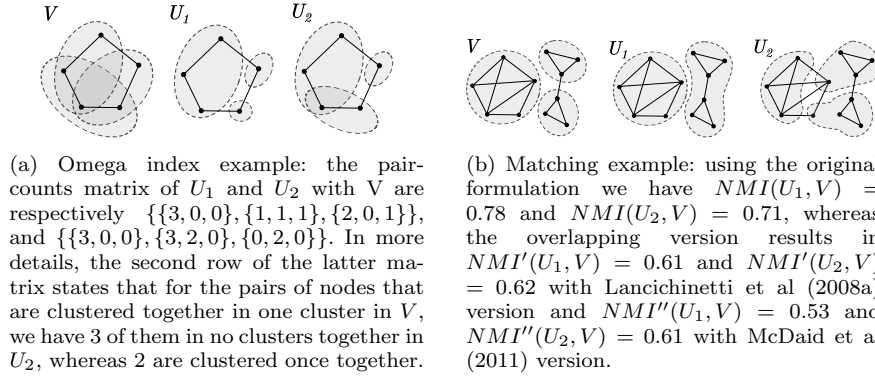


Fig. 2: Example for **Omega index** on the left: the pair-counts table for  $U_1$  and  $U_2$  with  $V$  have the same trace, and therefore they have the same degree of agreement with  $V$  according to the Omega index. Example for **the problem of matching** on the right: using the set matching based measures, such as the overlapping version of the  $NMI$ , clustering  $U_2$  is in higher agreement with  $V$ , while the non-overlapping version of  $NMI$  suggests the opposite. Here we used a disjoint example to be able to compare the results quantitatively with the original  $NMI$ , this problem is however intrinsic to all the matching based measures, regardless of the overlapping or disjoint.

Another commonly used measure for overlapping clusters (Gregory, 2010; Xie et al, 2013) is the extension of  $NMI$  proposed by Lancichinetti et al (2008a). The proposed measure does not reduce to the original  $NMI$  if the clusterings are disjoint. This extension assumes a matching between clusters in  $U$  and  $V$ , and only considers the best pair of clusters (with minimum conditional entropy) in the agreement calculation. A similar idea is also used in computing agreement between disjoint clusters, which is the basis of the set matching measures. These measures are known to suffer from the “problem of matching” (Meilă, 2007). See Figure 2b for a visualized example. The same problem exists with any of the agreement indexes that consider only the best matching, e.g. Balanced Error Rate with alignment, average  $F1$  score, and Recall measures used in (Yang and Leskovec, 2013; McAuley and Leskovec, 2012; McDaid et al, 2011). There is also a line of work on extensions for fuzzy clusters with soft membership (Brouwer, 2008; Quere et al, 2010; Campello, 2010; Anderson et al, 2010; Hullermeier et al, 2012). The fuzzy measures, however, are not applicable to cases where a data point could fully belong to more than one cluster, i.e. crisp overlapping (such as example of Figure 3) which are common in network clustering. The bonding concept presented by Brouwer (2008) is similar to the main idea behind our extension for overlapping cases, which we discuss further in Section 4.

The extension of the proposed  $\mathcal{D}$  formula (Definitions 1, 2, and 3) for overlapping clusters is not straightforward. The  $(\mathcal{A}/\mathcal{N})\mathcal{D}$  formula is indeed bounded for overlapping clusters, and reduces to the original formulation if we have disjoint covering clusters. However, the current formulation is not appropriate for comparing overlapping clusters, since it treats overlaps as variations and penalizes them. Consider an extreme example when we are comparing two identical clusterings, and therefore we should have  $(\mathcal{A}/\mathcal{N})\mathcal{D} = 0$  (i.e. the perfect agreement); this is true if there is no overlapping nodes, however as the number of overlapping nodes increase,  $(\mathcal{A}/\mathcal{N})\mathcal{D}$  also increase (i.e. the agreement decreases).

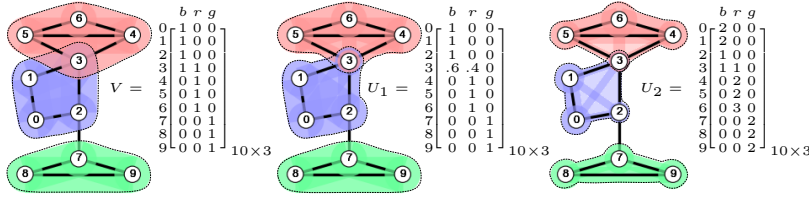


Fig. 3: Example of general matrix representation for a clustering:  $V$  and  $U_1$  are the classic overlapping clusters with crisp, and soft memberships respectively. Node 3 fully belongs to both blue and red clusters in  $V$ , wherein  $U_1$ , it belongs 60% to the blue cluster and 40% to the red cluster. This representation is general in a sense that it could encode membership of nodes to clusters in any form, with no assumptions on the matrix, such as in  $U_2$ .

The difficulty of computing the agreement of different clusterings, and in particular their extension for general cases such as overlapping clusters, partly comes from the fact that there is no matching between the clusters from the two clusterings. Therefore, one should consider all the permutations, or only consider the best matching, which is cursed with the “problem of matching” as discussed earlier. We overcome this difficulty by an alternative algebraic formulation for the clustering agreement measures, which takes the permutation out of the equation.

#### 4 Algebraic Formulation for Clustering Distance

Let  $U_{n \times k}$  denote a general representation for a clustering of a dataset with  $n$  datapoints, i.e.  $u_{ik}$  represents the memberships of node  $i$  in the  $k^{th}$  cluster of  $U$ . Different constraints on this representation derive different cases of clustering. For crisp clusters (a.k.a strict membership),  $u_{ik}$  is restricted to 0, 1 (1 if node  $i$  belongs to cluster  $k$  and 0 otherwise); whereas for probabilistic clusters (or soft membership),  $u_{ik}$  could be any real number in  $[0, 1]$ ; see Figure 3 for examples. Fuzzy clusters usually assume an additional constraint that the total membership of a datapoint is equal to one, i.e.  $u_{i.} = \sum_k u_{ik} = 1$ . Which should also be true for disjoint clusters, as each datapoint can only belong to one cluster.

Here we first show that the clustering agreement measures discussed before can be reformulated in terms of this matrix representation. The size of overlaps between clusters in  $U_{d \times k}$  and  $V_{d \times r}$ —their contingency matrix— derives as:

$$N = (U^T V)_{k \times r} = (V^T U)_{k \times r}^T$$

The agreement between disjoint clustering  $U$  and  $V$  is then calculated based on this contingency table. More specifically, we can reformulate  $\mathcal{D}$  and  $\mathcal{N}\mathcal{D}$  as:

$$\mathcal{D}_\varphi = \left[ \mathbf{1}\varphi(N\mathbf{1}^T) - \mathbf{1}\varphi(N)\mathbf{1}^T \right] + \left[ \varphi(\mathbf{1}N)\mathbf{1}^T - \mathbf{1}\varphi(N)\mathbf{1}^T \right], \quad \mathcal{N}\mathcal{D}_\varphi = \frac{\mathcal{D}_\varphi}{\varphi(\mathbf{1}N\mathbf{1}^T)}$$

where  $\mathbf{1}$  is a vector of ones with appropriate shape so that the matrix-vector product is valid, i.e.  $\mathbf{1}N = [n_{.1}, n_{.2}, \dots, n_{.r}]$ , and  $N\mathbf{1}^T = [n_{1.}, n_{2.}, \dots, n_{k.}]^T$ ; and  $\varphi$  is applied element-wise to the given matrix. We can show that similar to the Identity 1 and 2, the normalized Variation of Information derives from  $\varphi(x) = x \log x$ ; and with  $\varphi(x) = \binom{x}{2}$ ,  $1 - \mathcal{N}\mathcal{D}_\varphi$  is equivalent to the rand index.

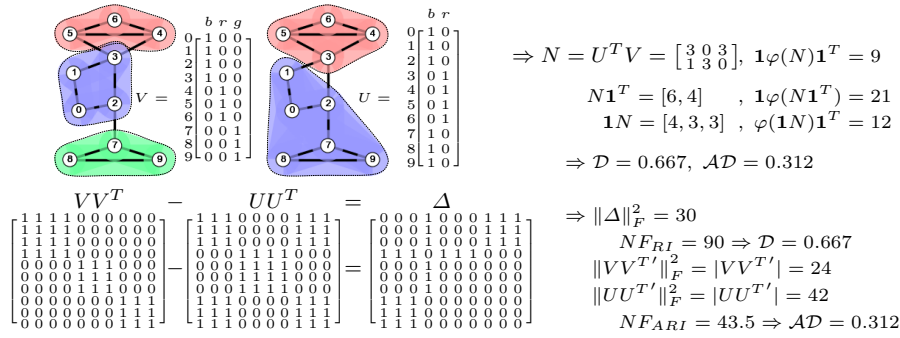


Fig. 4: Example for contingency v.s. co-membership based formulation. The  $(A)RI$  is first derived from the contingency table  $N$ , using  $\mathcal{D}$  formula where  $\varphi(x) = x(x-1)/2$ . Then same results are derived from the comparison of co-membership matrices  $UU^T$  and  $VV^T$ , using the alternative formulation of  $\mathcal{D}$ , where  $A'_{n \times n} = A - \mathbf{I}_n$  (see Footnote 3 for details).

Similarly,  $\mathcal{AD}$  can be reformulated as:

$$\mathcal{AD}_\varphi = \frac{\mathcal{D}_\varphi}{\frac{1}{2}[\mathbf{1}\varphi(N\mathbf{1}^T) + \varphi(\mathbf{1}N)\mathbf{1}^T] - E}, \quad E = \mathbf{1}\varphi\left(\frac{(N\mathbf{1}^T) \times (\mathbf{1}N)}{\mathbf{1}N\mathbf{1}^T}\right)\mathbf{1}^T$$

These formulations based contingency matrix of  $U^T V$ , as discussed in Section 3.2, are only appropriate for disjoint clusters. Therefore we propose the following reformulation of Definition 4, which is valid for both disjoint and overlapping cases. Instead of overlap matrix  $U^T V$ , Definition 4 measures the distance between clusterings directly from the difference of their co-membership matrices, i.e.  $UU^T - VV^T$ . This is inspired by the analogy between co-membership and overlap, i.e.  $(UU^T)_{ij}$  denotes in how many clusters node  $i$  and  $j$  appeared together, and  $(U^T U)_{ij}$  denotes how many nodes clusters  $i$  and  $j$  have in common.

**Definition 4** Co-Membership Clustering Difference ( $\Delta$ )

$$\Delta(U, V) = UU^T - VV^T, \quad \delta(U, V) = \frac{\Phi(\Delta)}{NF(U, V)}$$

where  $\Phi : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$  is a matrix function which quantifies the given difference matrix, e.g. a matrix norm, and  $NF(U, V)$  is a normalizing factor or an upper bound for  $\Phi(U, V)$ .

**Theorem 1** For disjoint clusters, the approximate<sup>3</sup>  $RI$  and  $ARI$  (Equation 6) derive from  $\Delta$  (proof in Appendix A.6), i.e.

$$\Phi = \|\cdot\|_F^2 \wedge NF = n^2 \times \max(\max(UU^T), \max(VV^T)) \Rightarrow 1 - \delta \equiv RI'(U, V)$$

$$\Phi = \|\cdot\|_F^2 \wedge NF = \|UU^T\|_F^2 + \|VV^T\|_F^2 - 2 \frac{|UU^T||VV^T|}{n^2} \Rightarrow 1 - \delta \equiv ARI'(U, V)$$

where  $|\cdot|$  is sum of all elements in the matrix, and  $\|\cdot\|_F^2$  is the sum of squared values, a.k.a. squared Frobenius norm.

<sup>3</sup> The exact formula derive if we change  $n^2$  by  $n(n-1)$  for the  $RI$ , and for the  $ARI$  (Equation 5) to also set the diagonal elements of the co-membership matrices to zero, i.e.  $UU^{T'} = UU^T - \mathbf{I}_n$ . Since the original  $(A)RI$  formula counts only the co-memberships of pairs of nodes  $(i, j)$  where  $i \neq j$ . The approximate version also considers the co-memberships for each single node with itself in different clusters, which is more suitable for overlapping cases.

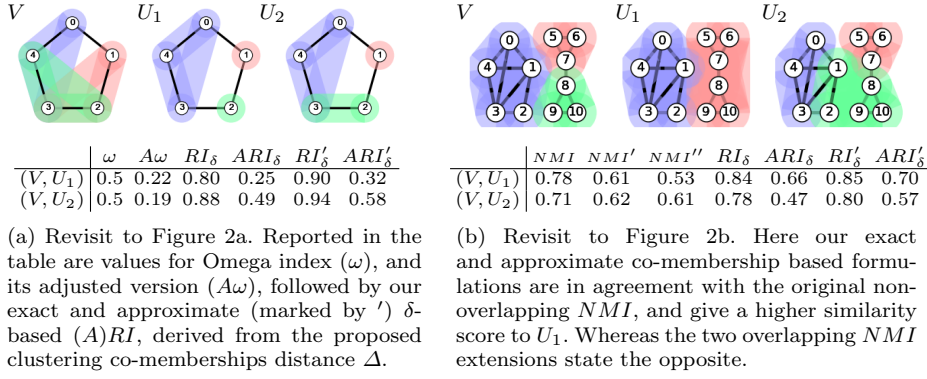


Fig. 5: Revisit to the examples of Figure 2. On the left we see that Omega index ( $\omega$ ) is unable to differentiate between  $U_1$  and  $U_2$ , whereas its adjusted version even gives higher score to  $U_1$ , which is the opposite of what we expect. The fact that  $U_2$  is more similar to  $V$  is captured by our  $\delta$ -based ( $A$ ) $RI$ . On the right we see an example of disagreement between the original  $NMI$  and its two set-matching based extensions for overlapping cases. Here since the problem is disjoint, ( $A$ ) $RI_\delta$  gives same results as the original ( $A$ ) $RI$ .

Our  $\delta$ -based formulation for  $RI$  and  $ARI$ , presented in Theorem 1, are also valid for overlapping cases. These formulations denoted respectively by  $RI_\delta$  and  $ARI_\delta$  hereafter, are identical to the original formulations if clusterings are disjoint; whereas unlike the overlap based formulations, they always return 1 if the clusterings are identical, regardless of the amount of the overlapping nodes. Refer to Appendix A.6 for more details, and see Figure 5 for examples.

It is worth mentioning that for crisp overlapping clusters, the Omega Index ( $\omega$ ) (Collins and Dent, 1988) derives from our formulation if we define  $\Delta = [UU^T == VV^T]$ , i.e.  $\Delta_{ij} = 1$  if  $(UU^T)_{ij} == (VV^T)_{ij}$  and zero otherwise. Then

$$\omega = |\Delta| - \text{tr}(\Delta), \quad A(\omega) = \frac{\omega - E[\omega]}{1 - E[\omega]}, \quad E[\omega] = \sum_{i=0}^{\min(r,k)} f_{UU^T}(i) f_{VV^T}(i)$$

where  $f_A(i)$  is the frequency of  $i$  in  $A$ . Figure 5a illustrates the effect of ignoring partial agreements by the  $\omega$  index. Similarly, we can compute other normalized forms of  $\Delta$ , or compare the co-membership matrices of  $UU^T$  and  $VV^T$  in other ways, e.g using matrix divergences (Dhillon and Tropp, 2007; Kulis et al, 2009). Here, we consider these two variations:

$$\mathcal{D}_{norm} = \frac{\|UU^T - VV^T\|_F^2}{\|UU^T\|_F^2 + \|VV^T\|_F^2}, \quad I_{\sqrt{tr}} = \frac{\text{tr}(UU^T VV^T)}{\sqrt{\text{tr}((UU^T)^2) \text{tr}((VV^T)^2)}} = \frac{|UU^T \circ VV^T|}{\|UU^T\|_F^2 \|VV^T\|_F^2}$$

It is also worth pointing out that in some applications, such as ensemble or multi-view clustering, we may not need the normalization and a measure of distance may suffice.

#### 4.1 Extension for Network Clustering

Here we define structure dependent clustering distances which incorporate the underlying structure of the graph. Let  $N$  denote the incidence matrix of the graph  $G$ , such that  $N_{ik} = \sqrt{A_{ij}}$  if node  $i$  is incident with edge  $k = (i, j)$ , and zero otherwise. Assuming a clustering as a transformation which assigns each datapoint

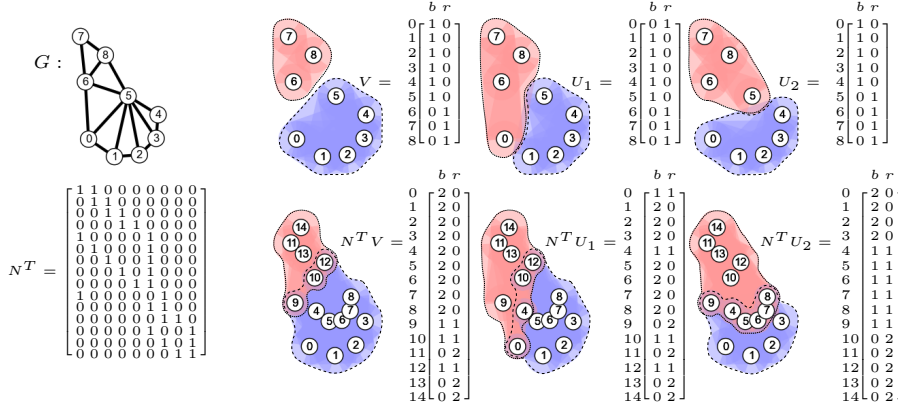


Fig. 6: A revisits to the example of Figure 1. Top) In the original data and considering only nodes,  $U_1$  and  $U_2$  have the same agreement with  $V$ . Since both  $U_1$  and  $U_2$  have one node clustered differently than  $V$ . Bottom) Transformed data using corresponding clusterings correctly identifies that  $U_1$  is closer to  $V$  compared to  $U_2$ . Note that the transformed data is similar to the line graph (edges as nodes) of the original data.

| -                 | $RI_\delta$ | $ARI_\delta$ | $RI'_\delta$ | $ARI'_\delta$ | $I_{norm}$ | $I_{\sqrt{lr}}$ |
|-------------------|-------------|--------------|--------------|---------------|------------|-----------------|
| $(U, V_1)$        | 0.778       | 0.556        | 0.802        | 0.604         | 0.695      | 0.815           |
| $(U, V_2)$        | 0.778       | 0.556        | 0.802        | 0.604         | 0.695      | 0.815           |
| $\perp(U, V_1 G)$ | 0.926       | 0.744        | 0.928        | 0.752         | 0.799      | 0.923           |
| $\perp(U, V_2 G)$ | 0.857       | 0.417        | 0.859        | 0.435         | 0.708      | 0.844           |
| $+(U, V_1 G)$     | 0.889       | 0.773        | 0.901        | 0.797         | 0.843      | 0.904           |
| $+(U, V_2 G)$     | 0.833       | 0.660        | 0.900        | 0.776         | 0.832      | 0.885           |
| $(N, U)$          | 0.750       | 0.500        | 0.979        | 0.327         | 0.512      | 0.662           |
| $(N, V_1)$        | 0.750       | 0.491        | 0.979        | 0.337         | 0.503      | 0.668           |
| $(N, V_2)$        | 0.639       | 0.264        | 0.977        | 0.275         | 0.481      | 0.616           |

Table 1: Results of different agreements for the example of Figure 1. The first two rows show that all the original structure independent measures result in the same agreement for  $U_1$  and  $U_2$ . Whereas the structure based measures give higher agreement score to  $U_1$  compared to  $U_2$ . The last three rows give the agreement of each clustering with the structure of the graph.

to one of its  $k$  clusters, i.e.  $U : n \mapsto k$ , we can incorporate the structure by measuring the distance between the transformed data by  $U$  and  $V$  as:

$$\mathcal{D}_\perp(U, V|G) = \mathcal{D}(N^T U, N^T V)$$

This is similar to measuring the structure similarity by counting the edges of the subgraphs, proposed earlier in Section 3.1; See Figure 6 for an example. We should note that the above formulation requires an overlapping distance, such as  $ARI_\delta$ .

Alternatively, we can assume each edge as a cluster of two nodes, and measure the distance of a clustering from the underlying structure of the graph. Consequently, the structure dependent distance of  $U$  and  $V$  can be defined as a combination of  $\mathcal{D}(U, N)$ ,  $\mathcal{D}(V, N)$  and  $\mathcal{D}(U, V)$ , for example:

$$\mathcal{D}_+(U, V|G) = \alpha \mathcal{D}(U, V) + (1 - \alpha) |\mathcal{D}(U, N) - \mathcal{D}(V, N)|, \quad \alpha = 0.5$$

Table 1, Table 2 and Table 3 compare structure dependent and independent measures for our earlier examples in Figure 1, and Figure 2. Wherein the experiments of the next section compare the measures in the context of community mining evaluation.

| -                 | $RI_\delta$ | $ARI_\delta$ | $RI'_\delta$ | $ARI'_\delta$ | $I_{norm}$ | $I_{\sqrt{tr}}$ |
|-------------------|-------------|--------------|--------------|---------------|------------|-----------------|
| $(V, U_1)$        | 0.800       | 0.245        | 0.902        | 0.318         | 0.532      | 0.764           |
| $(V, U_2)$        | 0.875       | 0.490        | 0.942        | 0.577         | 0.663      | 0.894           |
| $\perp(V, U_1 G)$ | 0.856       | 0.186        | 0.868        | 0.211         | 0.536      | 0.860           |
| $\perp(V, U_2 G)$ | 0.913       | 0.427        | 0.924        | 0.483         | 0.672      | 0.961           |
| $+(V, U_1 G)$     | 0.775       | 0.556        | 0.919        | 0.617         | 0.720      | 0.859           |
| $+(V, U_2 G)$     | 0.863       | 0.712        | 0.954        | 0.765         | 0.824      | 0.945           |
| $(N, U)$          | 0.850       | 0.333        | 0.933        | 0.528         | 0.682      | 0.816           |
| $(N, U_1)$        | 0.600       | 0.200        | 0.870        | 0.444         | 0.590      | 0.771           |
| $(N, U_2)$        | 0.700       | 0.400        | 0.900        | 0.576         | 0.666      | 0.822           |

Table 2: Results of different agreements for the omega example of Figure 2a.

| -                 | $RI_\delta$ | $ARI_\delta$ | $RI'_\delta$ | $ARI'_\delta$ | $I_{norm}$ | $I_{\sqrt{tr}}$ |
|-------------------|-------------|--------------|--------------|---------------|------------|-----------------|
| $(V, U_1)$        | 0.836       | 0.660        | 0.851        | 0.703         | 0.705      | 0.840           |
| $(V, U_2)$        | 0.782       | 0.471        | 0.802        | 0.567         | 0.626      | 0.721           |
| $\perp(V, U_1 G)$ | 0.900       | 0.790        | 0.906        | 0.806         | 0.768      | 0.902           |
| $\perp(V, U_2 G)$ | 0.857       | 0.564        | 0.862        | 0.607         | 0.667      | 0.798           |
| $+(V, U_1 G)$     | 0.855       | 0.708        | 0.922        | 0.793         | 0.839      | 0.866           |
| $+(V, U_2 G)$     | 0.818       | 0.556        | 0.897        | 0.716         | 0.782      | 0.804           |
| $(N, U)$          | 0.945       | 0.865        | 0.977        | 0.620         | 0.615      | 0.814           |
| $(N, U_1)$        | 0.818       | 0.621        | 0.970        | 0.502         | 0.589      | 0.707           |
| $(N, U_2)$        | 0.800       | 0.506        | 0.968        | 0.485         | 0.552      | 0.702           |

Table 3: Results of different agreements for the matching example of Figure 2b.

## 5 Experimental Results

Clustering agreement measures are often used in external evaluation of clustering algorithms, i.e. to compare their results with the known ground-truth in the benchmark datasets (Lancichinetti and Fortunato, 2009b). Here we perform similar sets of experiments, however the purpose is not to compare the general performance of community mining methods, but rather to show different comparisons/rankings we obtained using different agreement measures. Three sets of results are presented in the following to compare i) classic agreement indexes, ii) structure dependent and independent indexes, and iii) overlapping extensions.

### 5.1 Experiment Settings

In each experiment we select a set of common community mining methods, which discover clusters in a given network from different methodologies. In case of disjoint partitioning for Section 5.2 and 5.3, we use Louvain by Blondel et al (2008), Walk-Trap by Pons and Latapy (2005), PottsModel by Ronhovde and Nussinov (2009), FastModularity by Newman (2004), and InfoMap by Rosvall and Bergstrom (2008). For Section 5.4 we select four overlapping community detection methods: COPRA by Gregory (2010), MOSES by McDaid and Hurley (2010), OSLOM by Lancichinetti et al (2011), and BIGCLAM by Yang and Leskovec (2013). The authors' original implementations are used for all the algorithms, with no parameter tuning (defaults are used); and the reported agreements are averaged over ten runs.

Datasets are generated using the LFR (Lancichinetti et al, 2008b) benchmarks, which are commonly used in the evaluation of community mining algorithms. Parameters are chosen similar to the experiments by Lancichinetti and Fortunato

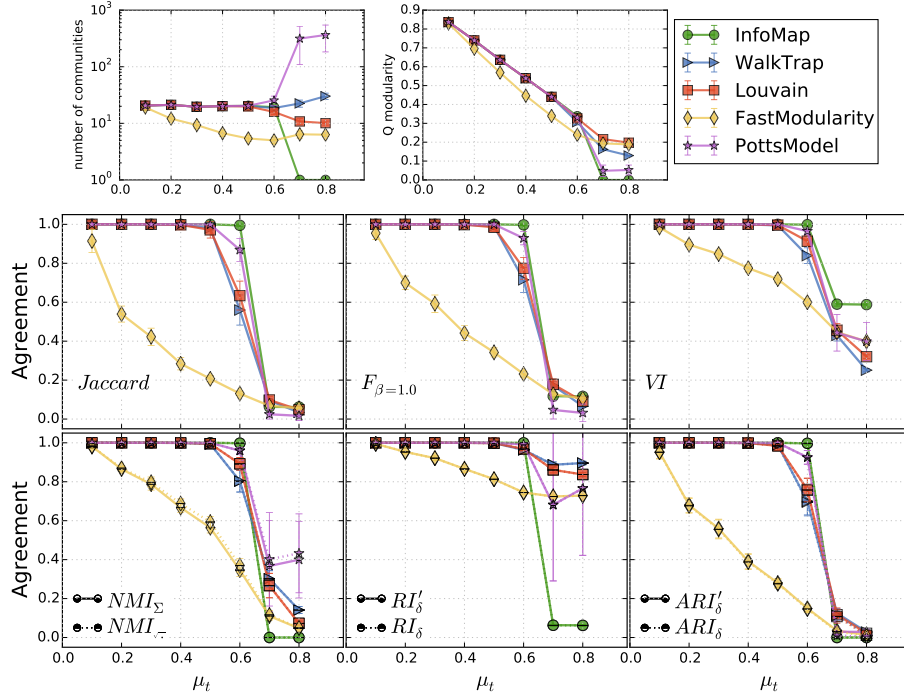


Fig. 7: The agreement of results from different community detection algorithms with the ground-truth in *unweighted LFR benchmarks*, plotted as a function of the mixing parameter. For large mixing parameters ( $\mu_t$ ),  $NMI_{\sqrt{\cdot}}$  and  $NMI_{\Sigma}$  rank PottsModel significantly higher since it finds too many communities; whereas  $VI$  (as opposite to  $RI$ ) marks InfoMap significantly better mainly because it resulted in too few communities; neither are close to the ground-truth. In the last three plots, similar measures are overlaid to show they are highly similar.

(2009b), i.e. networks with 1000 nodes, average degree of 20, max degree of 50, and power law degree exponent of -2; where the size of communities follows a power law distribution with exponent of -1, and ranges between 20 to 100 nodes. For the first experiments in Section 5.2, we generated *unweighted* LFR benchmarks with mixing parameters that varies from 0.1 to 0.8. Second experiment in Section 5.3 uses *weighted* LFR benchmarks (Lancichinetti and Fortunato, 2009a), where the mixing parameter for topology is fixed to 0.5, and the mixing parameter for weights varies. For the last experiment, we change the fraction of overlapping nodes, and generate unweighted LFR networks with the mixing parameter for topology fixed to 0.1, and 2 is set as the maximum number of communities a node can belong to, similar to experiments in (Lancichinetti et al, 2011). Results for other parameter settings, including smaller sized communities (10 to 50), could be found in the supplementary materials<sup>4</sup>.

<sup>4</sup> <https://github.com/rabbanyk/CommunityEvaluation>

## 5.2 Classic Measures

Figure 7 shows the comparison of the algorithms obtained by six different agreement measures<sup>5</sup>. Overall, the ranking of the algorithms according to these agreement measures is very similar. However, for large mixing parameters, the PottsModel is ranked significantly higher according to the  $NMI$  ( $NMI_\Sigma$  or  $NMI_{\sqrt{\cdot}}$ ), which is not consistent with the ranking obtained from the  $ARI$ , plotted as  $ARI_\delta$  in Figure 7. The  $\delta$  subscript indicates that the  $ARI$  is computed based on our  $\delta$ -based formulation, which is equivalent to the original  $ARI$  in this experiment, since communities are non-overlapping (Theorem 1). This disagreement for large mixing parameters is most probably because of the bias  $NMI$  has to the larger number of clusters (Vinh et al, 2009). Apart from this difference, the ranking from  $NMI$  is very similar to the one obtained from  $ARI$ . This is expected as these indices are measuring the same quantity as shown in the generalization of Definition 3. We can further see that there is no clear difference between the rankings from the approximate (See Footnote 3) and original  $ARI$ , i.e.  $ARI'_\delta$  and  $ARI_\delta$  in the Figure 7. This is desirable as we can use them interchangeably, whilst the former is more appropriate in the case of overlapping clusters, as discussed in Section 4.

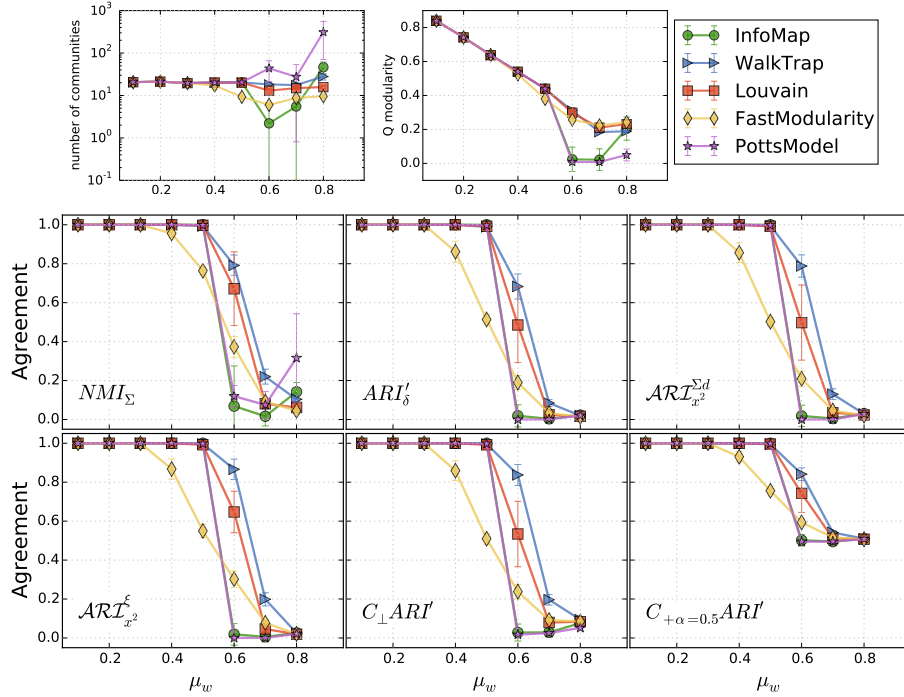


Fig. 8: Comparison of agreement indexes on *weighted LFR benchmark*, plotted as a function of the mixing parameter for weights. The difference between WalkTrap and Louvain is more significant according to the  $ART_{x^2}^\xi$ ,  $ART_{x^2}^{\Sigma^d}$ , and  $C_\perp ARI'$  which are structure dependent measures. We can also see the bias of  $NMI$  to the number of clusters, similar to the Figure 7.

<sup>5</sup> Similar trends are observed for other variations of agreement measures which can be found in the supplementary materials.



### 5.3 Structure Dependent Measures

Figure 8 compares the community mining methods over the *weighted* LFR benchmarks. Similar to the previous experiment, the rankings are very close. However, the difference between structure dependent and independent measures has become clear with the presence of weights. We can see that the Walktrap method is performing better according to most of the measures, whereas the distinction is more readable in the structure dependent variations: i.e. the degree weighted ARI ( $\mathcal{ARI}_{x_2}^\xi$ ), and edge counting ARI ( $\mathcal{ARI}_{x_2}^{\Sigma d}$ ) introduced in Section 3.1, and the transformed ARI ( $C_\perp \text{ARI}'$ ) introduced in Section 4.1.

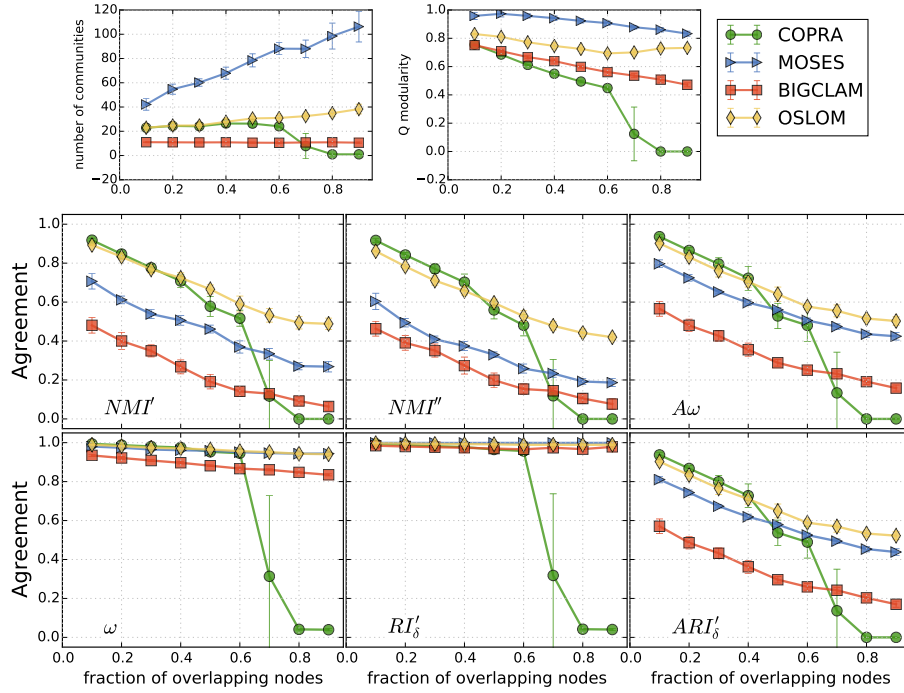


Fig. 9: Comparison of agreement indexes on *overlapping LFR benchmark*, plotted as a function of the fraction of overlapping nodes. We can see the negative bias of number of clusters in the set matching overlapping measures, i.e.  $NMI'$  and  $NMI''$ , which strongly penalize MOSES for finding many communities. Which is not the case with  $A\omega$  and  $ARI'_\delta$ . We can also see the impracticability of un-adjusted measures:  $\omega$  and  $RI'_\delta$ , which are also very similar.

### 5.4 Overlapping Measures

Figure 9 shows the comparison of these methods based on different overlapping agreement indexes: the overlapping NMI variations:  $NMI'$  by Lancichinetti et al (2008a) and  $NMI''$  from McDaid et al (2011); the omega index ( $\omega$ ), and its adjusted version ( $A\omega$ ); and our  $\delta$ -based formulations for the  $RI$  and  $ARI$ , i.e.  $RI'_\delta$ , and  $ARI'_\delta$ . Here also we observe a generally similar ranking. However the difference

between MOSES and OSLOM is more significant according to the set-matching based extensions of *NMI*. This most probably is because MOSES finds much more communities, and hence it is more likely for it to have communities that do not get matched/compared with the communities in the ground-truth, although they show valid groupings of the nodes. We can also see that in this case, the ranking from adjusted omega,  $A\omega$  and  $ARI'_\delta$  are very similar, which can be explained as in our settings, each node can only belong to maximum of two communities; whereas the difference between  $A\omega$  and  $ARI'_\delta$  becomes clear if a node can belong to many communities.

## 6 Conclusion

In this paper, we presented generalizations of clustering agreement measures. This generalization illustrates the relation between the Rand Index (*RI*) and Variation of Information (*VI*); and Adjusted Rand Index (*ARI*) and the Normalized Mutual Information (*NMI*). We then discussed the necessity of structure dependent agreement measures, particularly in the evaluation of clusters over networks, i.e. communities; and proposed extensions of the general formula for such cases. We further discussed the difficulty of extending this contingency based formula for overlapping clusters and proposed reformulation which works for overlapping cases. We showed that the original *RI* and *ARI* of non-overlapping clusters derive from this reformulation.

## References

- Aggarwal CC, Reddy CK (2014) Data Clustering: Algorithms and Applications. CRC Press
- Albatineh AN, Niewiadomska-Bugaj M, Mihalko D (2006) On similarity indices and correction for chance agreement. *Journal of Classification* 23:301–313, 10.1007/s00357-006-0017-z
- Anderson DT, Bezdek JC, Popescu M, Keller JM (2010) Comparing Fuzzy, Probabilistic, and Possibilistic Partitions. *IEEE Transactions on Fuzzy Systems* 18(5):906–918
- Banerjee A, Merugu S, Dhillon IS, Ghosh J (2005) Clustering with Bregman Divergences. *The Journal of Machine Learning Research* 6:1705–1749
- Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008:P10,008+
- Brouwer RK (2008) Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. *Journal of Intelligent Information Systems* 32(3):213–235
- Campello RR (2010) Generalized external indexes for comparing data partitions with overlapping categories. *Pattern Recognition Letters* 31(9):966–975
- Collins LM, Dent CW (1988) Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. *Multivariate Behavioral Research* 23(2):231–242
- Cortina-Borja M (2012) Handbook of parametric and nonparametric statistical procedures, 5th edn. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175(3):829–829

- Cui Y, Fern X, Dy J (2007) Non-redundant multi-view clustering via orthogonalization. *Data Mining, 2007 ICDM 2007* ...
- Dhillon IS, Tropp JA (2007) Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications* 29(4):1120–1146
- Fortunato S (2010) Community detection in graphs. *Physics Reports* 486(35):75–174
- Gregory S (2010) Finding overlapping communities in networks by label propagation. *New Journal of Physics* 12(10):103,018
- Hubert L, Arabie P (1985) Comparing partitions. *Journal of Classification* 2:193–218
- Hullermeier E, Rifqi M, Henzgen S, Senge R (2012) Comparing Fuzzy Partitions: A Generalization of the Rand Index and Related Measures. *IEEE Transactions on Fuzzy Systems* 20(3):546–556
- Kulis B, Sustik MA, Dhillon IS (2009) Low-rank kernel learning with bregman matrix divergences. *The Journal of Machine Learning Research* 10:341–376
- Lancichinetti A, Fortunato S (2009a) Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* 80(1):016,118
- Lancichinetti A, Fortunato S (2009b) Community detection algorithms: A comparative analysis. *Physical Review E* 80(5):056,117
- Lancichinetti A, Fortunato S, Kertesz J (2008a) Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics* 11(3):20
- Lancichinetti A, Fortunato S, Radicchi F (2008b) Benchmark graphs for testing community detection algorithms. *Physical Review E* 78(4):046,110
- Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. *PloS one* 6(4):e18,961
- Light RJ, Margolin BH (1971) An analysis of variance for categorical data. *Journal of the American Statistical Association* 66(335):pp. 534–544
- Manning CD, Raghavan P, Schtze H (2008) *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA
- McAuley J, Leskovec J (2012) Discovering social circles in ego networks. *arXiv preprint arXiv:12108182*
- McDaid A, Hurley N (2010) Detecting highly overlapping communities with model-based overlapping seed expansion. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, IEEE, pp 112–119
- McDaid AF, Greene D, Hurley N (2011) Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:11102515*
- Meilă M (2007) Comparing clusteringsan information based distance. *Journal of Multivariate Analysis* 98(5):873–895
- Newman ME (2004) Fast algorithm for detecting community structure in networks. *Physical review E* 69(6):066,133
- Nielsen F, Nock R (2013) On the Chi square and higher-order Chi distances for approximating f-divergences p 11, 1309.3029
- Pons P, Latapy M (2005) Computing communities in large networks using random walks. In: *Computer and Information Sciences-ISCIS 2005*, Springer, pp 284–293
- Quere R, Le Capitaine H, Fraisseix N, Frelicot C (2010) On Normalizing Fuzzy Coincidence Matrices to Compare Fuzzy and/or Possibilistic Partitions with the

- Rand Index. In: 2010 IEEE International Conference on Data Mining, IEEE, pp 977–982
- Ronhovde P, Nussinov Z (2009) Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E* 80(1):016,109
- Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105(4):1118–1123
- Vinh NX, Epps J, Bailey J (2009) Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, New York, NY, USA, ICML '09, pp 1073–1080
- Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research* 11:2837–2854
- Warrens MJ (2008a) On similarity coefficients for 22 tables and correction for chance. *Psychometrika* 73:487–502, 10.1007/s11336-008-9059-y
- Warrens MJ (2008b) On the Equivalence of Cohen’s Kappa and the Hubert-Arabie Adjusted Rand Index. *Journal of Classification* 25:177–183
- Wu J, Xiong H, Chen J (2009) Adapting the right measures for k-means clustering. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, KDD '09, pp 877–886
- Xie J, Kelley S, Szymanski BK (2013) Overlapping community detection in networks. *ACM Computing Surveys* 45(4):1–35, 1110.5813
- Yang J, Leskovec J (2013) Overlapping community detection at scale: a nonnegative matrix factorization approach. In: *Proceedings of the sixth ACM international conference on Web search and data mining*, ACM, pp 587–596
- Zhou D, Li J, Zha H (2005) A new mallows distance based metric for comparing clusterings. In: *Proceedings of the 22nd international conference on Machine learning*, ACM, pp 1028–1035



## A Proofs

**A.1 Proof of Proposition 1:** From the definition of Variation of information we have:

$$VI(U, V) = H(U) + H(V) - 2I(U, V) = 2H(U, V) - H(U) - H(V) = \mathbf{H}(\mathbf{V}|\mathbf{U}) + \mathbf{H}(\mathbf{U}|\mathbf{V})$$

On the other hand, we have:

$$\begin{aligned} RI(U, V) &\propto \frac{1}{n^2 - n} \left( \sum_{i=1}^k \left[ \sum_{j=1}^r n_{ij}^2 - \left( \sum_{j=1}^r n_{ij} \right)^2 \right] + \sum_{j=1}^r \left[ \sum_{i=1}^k n_{ij}^2 - \left( \sum_{i=1}^k n_{ij} \right)^2 \right] \right) \\ &\stackrel{*}{\propto} \sum_{i=1}^k [E_j(n_{ij}^2) - E_j(n_{ij})^2] + \sum_{j=1}^r [E_i(n_{ij}^2) - E_i(n_{ij})^2] \\ &\stackrel{*}{\propto} \sum_{i=1}^k Var_j(n_{ij}) + \sum_{j=1}^r Var_i(n_{ij}) \stackrel{**}{\propto} \mathbf{Var}(\mathbf{V}|\mathbf{U}) + \mathbf{Var}(\mathbf{U}|\mathbf{V}) \quad \square \end{aligned}$$

(\*)  $E_j/Var_j$  shows the average/variance of values in the  $j^{th}$  column of the contingency table.

(\*\*) The RI is in fact proportional to the average variance of rows/columns values in the contingency table, which we denote by conditional variance. For other forms of conditional variance for categorical data see Light and Margolin (1971).

**A.2 Proof of Corollary 1:** We first show that  $0 \leq \mathcal{D}_\varphi^\eta(U||V)$  which also results in the lower bound 0 for  $\mathcal{D}_\varphi^\eta(U, V)$  since,  $\mathcal{D}_\varphi^\eta(U, V) = \mathcal{D}_\varphi^\eta(U||V) + \mathcal{D}_\varphi^\eta(V||U)$ . From the superadditivity of  $\varphi$  we have:

$$\sum_{u \in U} \varphi(\eta_{uv}) \leq \varphi\left(\sum_{u \in U} \eta_{uv}\right) \implies \sum_{v \in V} \left[ \varphi\left(\sum_{u \in U} \eta_{uv}\right) - \sum_{u \in U} \varphi(\eta_{uv}) \right] \geq 0 \implies \mathcal{D}_\varphi^\eta(\mathbf{U}||\mathbf{V}) \geq 0$$

Similarly for the upper bound, from positivity and super-additivity we get respectively:

$$\mathcal{D}_\varphi^\eta(U||V) = \sum_{v \in V} \varphi\left(\sum_{u \in U} \eta_{uv}\right) - \sum_{v \in V} \sum_{u \in U} \varphi(\eta_{uv}) \leq \sum_{v \in V} \varphi\left(\sum_{u \in U} \eta_{uv}\right) \leq \varphi\left(\sum_{v \in V} \sum_{u \in U} \eta_{uv}\right)$$

**A.3 Proof of Identity 1:** The proof is elementary, if we write the definition for  $\varphi = x \log x$ , we get:

$$\begin{aligned} \mathcal{N}\mathcal{D}_{x \log x}^{|\cap|}(U, V) &= \frac{\sum_{v \in V} \sum_{u \in U} |u \cap v| [\log(\sum_{u \in U} |u \cap v|) - \log(|u \cap v|)]}{(\sum_{v \in V} \sum_{u \in U} |u \cap v|) \log(\sum_{v \in V} \sum_{u \in U} |u \cap v|)} \\ &+ \frac{\sum_{u \in U} \sum_{v \in V} |u \cap v| [\log(\sum_{v \in V} |u \cap v|) - \log(|u \cap v|)]}{(\sum_{u \in U} \sum_{v \in V} |u \cap v|) \log(\sum_{u \in U} \sum_{v \in V} |u \cap v|)} \\ &\stackrel{*}{=} \frac{\sum_j^r \sum_i^k n_{ij} [\log(\sum_i^k n_{ij}) + \log(\sum_j^r n_{ij}) - 2 \log(n_{ij})]}{(\sum_i^k \sum_j^r n_{ij}) \log(\sum_i^k \sum_j^r n_{ij})} \\ &\stackrel{**}{=} \frac{1}{\log n} \sum_j^r \sum_i^k \frac{n_{ij}}{n} \log\left(\frac{n_{i.} n_{.j}}{n_{ij}^2}\right) = \frac{VI(U, V)}{\log n} \quad \square \end{aligned}$$

(\*) slight change of notation, i.e. from  $\sum_{u \in U}$  to  $\sum_i^k$ ,  $\sum_{v \in V}$  to  $\sum_j^r$  and  $|u \cap v|$  to  $n_{ij}$ .

(\*\*) since  $\sum_i^k \sum_j^r n_{ij} = n$ ,  $\sum_i^k n_{ij} = n_{.j}$  and  $\sum_j^r n_{ij} = n_{i.}$ .

**A.4 Proof of Identity 2:** Similar to the previous proof from the definition we derive:

$$\begin{aligned} \mathcal{N}\mathcal{D}_{\binom{x}{2}}^{|\cap|}(U, V) &\stackrel{*}{=} \frac{\sum_j^r \left[ (\sum_i^k n_{ij})^2 - \sum_i^k n_{ij}^2 \right] + \sum_i^k \left[ (\sum_j^r n_{ij})^2 - \sum_j^r n_{ij}^2 \right]}{(\sum_i^k \sum_j^r n_{ij})^2 - \sum_i^k \sum_j^r n_{ij}^2} \\ &\stackrel{**}{=} \frac{1}{n^2 - n} \left[ \sum_j^r (n_{.j})^2 + \sum_i^k (n_{i.})^2 - 2 \sum_j^r \sum_i^k n_{ij}^2 \right] = 1 - RI(U, V) \quad \square \end{aligned}$$

(\*), (\*\*) same as previous proof.

### A.5 Proof of Identity 3 and 4:

$$\begin{aligned}\mathcal{AD}_\varphi^\eta &= \frac{\sum_{v \in V} \varphi(\eta.v) + \sum_{u \in U} \varphi(\eta.u.) - 2 \sum_{v \in V} \sum_{u \in U} \varphi(\eta_{uv})}{\sum_{v \in V} \varphi(\eta.v) + \sum_{u \in U} \varphi(\eta.u.) - 2 \sum_{u \in U} \sum_{v \in V} \varphi\left(\frac{\eta.v \eta.u.}{\sum_{u \in U} \sum_{v \in V} \eta_{uv}}\right)} \\ \Rightarrow 1 - \mathcal{AD}_\varphi^\eta(U, V) &= \frac{\sum_{v \in V} \sum_{u \in U} \varphi(\eta_{uv}) - \sum_{u \in U} \sum_{v \in V} \varphi\left(\frac{\eta.v \eta.u.}{\sum_{u \in U} \sum_{v \in V} \eta_{uv}}\right)}{\frac{1}{2} [\sum_{v \in V} \varphi(\eta.v) + \sum_{u \in U} \varphi(\eta.u.)] - \sum_{u \in U} \sum_{v \in V} \varphi\left(\frac{\eta.v \eta.u.}{\sum_{u \in U} \sum_{v \in V} \eta_{uv}}\right)}\end{aligned}$$

This formula resembles the adjustment for chance in Equation 4, where the measure being adjusted is  $\sum_{v \in V} \sum_{u \in U} \varphi(\eta_{uv})$ , the upper bound used for it is  $\frac{1}{2} [\sum_{v \in V} \varphi(\eta.v) + \sum_{u \in U} \varphi(\eta.u.)]$ , and the expectation is defined as:

$$E[\sum_{v \in V} \sum_{u \in U} \varphi(\eta_{uv})] = \sum_{u \in U} \sum_{v \in V} \varphi\left(\frac{\eta.v \eta.u.}{\sum_{u \in U} \sum_{v \in V} \eta_{uv}}\right)$$

Now if we have  $\varphi(xy) = \varphi(x)\varphi(y)$ , which is true for  $\varphi(x) = x^2$ , we have:

$$E[\sum_{v \in V} \sum_{u \in U} \varphi(\eta_{uv})] = \sum_{u \in U} \sum_{v \in V} \frac{\varphi(\eta.v) \varphi(\eta.u.)}{\varphi(\sum_{u \in U} \sum_{v \in V} \eta_{uv})} = \frac{\sum_{v \in V} \varphi(\eta.v) \sum_{u \in U} \varphi(\eta.u.)}{\varphi(\sum_{u \in U} \sum_{v \in V} \eta_{uv})}$$

Using this expectation, if we substitute  $\varphi = x^2$  we get the  $ARI'$  of Equation 6, and using the  $\varphi = \binom{x}{2}$  and the later reformulation of  $E$ , we get the original  $ARI$  of Equation 5, as:

$$\begin{aligned}1 - \mathcal{AD}_{\binom{x}{2}}^{|\cap|}(U, V) &= \frac{\sum_{v \in V} \sum_{u \in U} \binom{|u \cap v|}{2} - E(\sum_{v \in V} \sum_{u \in U} \binom{|u \cap v|}{2})}{\frac{1}{2} \left[ \sum_{v \in V} \binom{\sum_{u \in U} |u \cap v|}{2} + \sum_{u \in U} \binom{\sum_{v \in V} |u \cap v|}{2} \right] - E(\sum_{v \in V} \sum_{u \in U} \binom{|u \cap v|}{2})} \\ \text{where } E(\sum_{v \in V} \sum_{u \in U} \binom{|u \cap v|}{2}) &= \frac{\sum_{v \in V} \binom{\sum_{u \in U} |u \cap v|}{2} \sum_{u \in U} \binom{\sum_{v \in V} |u \cap v|}{2}}{\binom{n}{2}} \\ \Rightarrow 1 - \mathcal{AD}_{\binom{x}{2}}^{|\cap|}(U, V) &\stackrel{*,**}{=} \frac{\sum_j^r \sum_i^k \binom{n_{ij}}{2} - \sum_j^r \binom{n_{.j}}{2} \sum_i^k \binom{n_{i.}}{2} / \binom{n}{2}}{\frac{1}{2} \left[ \sum_j^r \binom{n_{.j}}{2} + \sum_i^k \binom{n_{i.}}{2} \right] - \sum_j^r \binom{n_{.j}}{2} \sum_i^k \binom{n_{i.}}{2} / \binom{n}{2}} = ARI(U, V) \quad \square \\ &(*), (**) \text{ same as proof of identity 1.}\end{aligned}$$

On the other hand for the  $NMI$ , we have:

$$\begin{aligned}1 - \mathcal{AD}_{x \log x}^{|\cap|}(U, V) &= \frac{\sum_{v \in V} \sum_{u \in U} n_{uv} \log n_{uv} - E(\sum_{v \in V} \sum_{u \in U} n_{uv} \log n_{uv})}{\frac{1}{2} \left[ \sum_{v \in V} n_{.v} \log n_{.v} + \sum_{u \in U} n_{u.} \log n_{u.} \right] - E(\sum_{v \in V} \sum_{u \in U} n_{uv} \log n_{uv})} \\ \text{where } E(\sum_{v \in V} \sum_{u \in U} n_{uv} \log n_{uv}) &= \sum_{u \in U} \sum_{v \in V} \left( \frac{\eta.v \eta.u.}{\sum_{u \in U} \sum_{v \in V} \eta_{uv}} \right) \log \left( \frac{\eta.v \eta.u.}{\sum_{u \in U} \sum_{v \in V} \eta_{uv}} \right) \\ \Rightarrow 1 - \mathcal{AD}_{x \log x}^{|\cap|}(U, V) &\stackrel{*,**}{=} \frac{\sum_j^r \sum_i^k n_{ij} \log n_{ij} - \sum_i^k \sum_j^r \frac{n_{.j} n_{i.}}{n} \log \frac{n_{.j} n_{i.}}{n}}{\frac{1}{2} \left[ \sum_j^r n_{.j} \log n_{.j} + \sum_i^k n_{i.} \log n_{i.} \right] - \sum_i^k \sum_j^r \frac{n_{.j} n_{i.}}{n} \log \frac{n_{.j} n_{i.}}{n}} \\ &= \frac{n \sum_j^r \sum_i^k \frac{n_{ij}}{n} \log \frac{n_{ij}}{n} + n \log n - \sum_i^k \sum_j^r \frac{n_{.j} n_{i.}}{n} [\log \frac{n_{.j}}{n} + \log \frac{n_{i.}}{n} + \log n]}{\frac{n}{2} \left[ \sum_j^r \frac{n_{.j}}{n} \log \frac{n_{.j}}{n} + \sum_i^k \frac{n_{i.}}{n} \log \frac{n_{i.}}{n} + 2 \log n \right] - \sum_i^k \sum_j^r \frac{n_{.j} n_{i.}}{n} [\log \frac{n_{.j}}{n} + \log \frac{n_{i.}}{n} + \log n]} \\ &= \frac{-H(U, V) + \log n - \sum_i^k \frac{n_{i.}}{n} \sum_j^r \frac{n_{.j}}{n} \log \frac{n_{.j}}{n} + \sum_i^k \frac{n_{i.}}{n} - \sum_j^r \frac{n_{.j}}{n} \log \frac{n_{.j}}{n} - \sum_i^k \sum_j^r \frac{n_{.j} n_{i.}}{n^2} \log n}{\frac{1}{2} [-H(U) - H(V)] + \log n + \sum_i^k \frac{n_{i.}}{n} H(V) + \sum_i^k \frac{n_{i.}}{n} H(U) - \log n} \\ &= \frac{-H(U, V) + H(V) + H(U)}{-\frac{1}{2} [H(U) + H(V)] + H(V) + H(U)} = \frac{I(U, V)}{\frac{1}{2} [H(U) + H(V)]} = NMI_{sum}(U, V) \quad \square \\ &(*), (**) \text{ same as proof of identity 1.}\end{aligned}$$

**A.6 Proof of Theorem 1:** First we prove that in general cases we have:

$$\|UU^T - VV^T\|_F^2 = \|U^T U\|_F^2 + \|V^T V\|_F^2 - 2\|U^T V\|_F^2$$

where  $\|\cdot\|_F^2$  is squared Frob norm. This holds since we have:

$$\begin{aligned} \|UU^T - VV^T\|_F^2 &= \sum_{ij} (UU^T - VV^T)_{ij}^2 \\ &= \sum_{ij} (UU^T)_{ij}^2 + \sum_{ij} (VV^T)_{ij}^2 - 2 \sum_{ij} (UU^T)_{ij} (VV^T)_{ij} \\ &= \|UU^T\|_F^2 + \|VV^T\|_F^2 - 2|UU^T \circ VV^T| \end{aligned}$$

Where the  $\circ$  is element-wise matrix product, a.k.a. hadamard product, and  $|\cdot|$  is sum of all elements in the matrix<sup>6</sup>. The proof is complete with showing:

$$\begin{aligned} |UU^T \circ VV^T| &= \text{tr}((UU^T)^T VV^T) = \text{tr}(V^T UU^T V) = \text{tr}((U^T V)^T U^T V) = \|U^T V\|_F^2 \\ \|UU^T\|_F^2 &= \text{tr}((UU^T)^T UU^T) = \text{tr}(U^T UU^T U) = \text{tr}((U^T U)^T U^T U) = \|U^T U\|_F^2 \end{aligned}$$

Now, we can prove Theorem 1 for the cases of disjoint hard clusters, using the notation,  $n_{ij} = (U^T V)_{ij}$ , we have  $\|U^T V\|_F^2 = \sum_{ij} n_{ij}^2$  and:

$$\|U^T U\|_F^2 = \sum_{ij} \langle U_{\cdot i}, U_{\cdot j} \rangle^2 = \sum_{ij} \left( \sum_k u_{ki} u_{kj} \right)^2 \stackrel{*}{=} \sum_i \left( \sum_k u_{ki}^2 \right)^2 \stackrel{**}{=} \sum_i \left( \sum_k u_{ki} \right)^2 \stackrel{***}{=} \sum_i n_i^2$$

(\*) with assumption that clusters are disjoint,  $u_{ki} u_{kj}$  is only non-zero iff  $i = j$

(\*\*) with the assumption that memberships are hard,  $u_{ki}$  is either 0 or 1, therefore  $u_{ki} = u_{ki}^2$

(\*\*\*) marginals of  $N$  give cluster sizes in  $U$  and  $V$ , i.e.  $n_i = \sum_j n_{ij} = \sum_k u_{ki} = |V_i|$

Therefore for disjoint hard clusters we get:

$$\|UU^T - VV^T\|_F^2 = \sum_i n_i^2 + \sum_j n_j^2 - 2 \sum_{ij} n_{ij}^2$$

The  $RI$  normalization assumes that all pairs are in disagreement, i.e.  $NF_{RI} = |\mathbf{1}_{n \times n}| = n^2$ , as  $\max(\max(UU^T), \max(VV^T)) = 1$ . The  $ARI$  normalization compares  $\Delta$  to the difference where the two random variable of  $UU_{ij}^T$  and  $VV_{ij}^T$  are independent, in which case we would have:

$$E(UU_{ij}^T VV_{ij}^T) = E((UU^T)_{ij}) E((VV^T)_{ij})$$

which is calculated by:

$$\frac{\sum_{ij} ((UU^T)_{ij} (VV^T)_{ij})}{n^2} = \frac{\sum_{ij} (UU^T)_{ij}}{n^2} \frac{\sum_{ij} (VV^T)_{ij}}{n^2}$$

Since  $\Delta = \|UU^T - VV^T\|_F^2 = \|UU^T\|_F^2 + \|VV^T\|_F^2 - 2\text{Sum}(UU^T \circ VV^T)$ , we have  $ARI = 0$  or  $\Delta/NF_{ARI} = 1$ , i.e. agreement no better than chance, when this independence condition holds, i.e.:

$$\Delta = NF_{ARI} \iff \text{Sum}(UU^T \circ VV^T) = \frac{|UU^T| |VV^T|}{n^2}$$

<sup>6</sup> This equality is also useful in the implementation to improve the scalability.